COMPUTER SOFTWARE PROGRAM FOR GRAPHICALLY DISPLAYING GENETIC LINKAGE UNBALANCES, AND THE METHOD THEREOF

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims priority under 35 U.S.C. 119 based upon Japanese Patent Application Serial No. 2003-48216, filed on January 21, 2003. The entire disclosure of the aforesaid applications is incorporated herein by reference.

BACKGROUND OF THE INVENTION

1. Field of the Invention

本発明は遺伝子多様性データ解析に関し、ケースデータ群とコントロールデータ群のそれぞれで求められたペアワイズ連鎖不平衡値を表示する上で、ケースデータの処理結果と、コントロールデータ群での処理結果を見やすく比較表示するための方法に関する。

2. Description of the Related Art

遺伝子多様性研究では、各遺伝子座位相互の連鎖の強さを計算することが頻繁に行われる。連鎖とはある遺伝子座位の多型と別に着目する座位の遺伝子多型がペアで子孫に遺伝していることを意味している。もし染色体上で十分離れていれば、遺伝子のランダムな組み換えが起こるため、５，６世代経過後にはほぼ平衡状態に落ち着くことが知られている。この状態をハーディワイバーグ平衡と呼ぶ。注目する遺伝子多様性の座位が物理的に近い場合、このハーディワインバーグ平衡からずれが保存される。このずれを連鎖不平衡と呼ぶ。

連鎖不平衡は２箇所のハプロタイプ度数情報を用いて、２×２の分割表を作成し、各座位でのハプロタイプ頻度から想定される独立の場合からどれだけずれているかを連鎖不平衡値として用いる。

まず、第一の遺伝子座位と第二の遺伝子座位のメジャーアレルを１、

マイナーアレルを３としてそれぞれのハプロタイプ頻度を以下のように表す。

第一遺伝子座位-第二遺伝子座位	頻度
1-1	p11
1-3	p13
3-1	p31
3-3	p33

ただし、p11, p13, p31, p33 は０から１の間の値で、p11 + p13 + p31 + p33 = 1 である。

すると、連鎖不平衡 D は次式で与えられる。

D=p11p33-p13p31

D は正負の値をとるが、0-1 間の値をとるように補正した D’という連鎖不平衡値も定義されている。D’は D>0orD=0 の場合は、D の取り得る最大値は、次式で与えられる。

Dmax=min(p1△×p△3,p3△×p△1)

ただし、p1△は第一座位のメジャーアレル頻度(p1△=p11+p13)、p△3は第二座位のマイナーアレル頻度(p△3=p13+p33)、同様に p3△は第一座位のマイナーアレル頻度(p3△=p31+p33)、p△1 は第二座位のメジャーアレル頻度(p△1=p11+p31)を意味する。

D<0 の場合は D の取り得る最小値は次式で与えられる。

Dmin=max(-p1△×p△1,-p3△×p△3)

これらを用い、

D'=D/Dmax(D が正の場合)

D'=D/Dmin(D が負の場合)

と定義される。

また、他に r^2 と呼ばれる連鎖不平衡値があり、次式で表される。

$r^2=D^2/(p1△×p3△×p△1×p△3)$

そのほかにも赤池情報量基準（以下 AIC と呼ぶ。Akaike's Information Criterion の略）を用いた方法などがある（K.Shimo-onodaetal : Akaike's

information criterion for a measure of linkaged isequilibrium, Journal of Human Genetics, Vol.47 Issue 12 (2002) pp649-655)。

これらの連鎖不平衡を表す指標値をケースデータ群とコントロールデータ群に対して求めることで疾患などのケース特有の連鎖不平衡の違いを持つ部分を見つけることが可能となる。

しかしながら、従来の技術では、単に連鎖不平衡の指標値を表形式で別々に表示しているだけであり、ケース・コントロール間での相違箇所を見つけ出すのが大変であるという問題があった。また、一塩基多型の検査データは数十個から多くは数千個以上まで対象とするため、全体的に見ながら相違点を見つけだすのが難しいという問題があった。

SUMMARY OF THE INVENTION

本発明は、上記のような課題を解決するために成されたもので、その目的は、遺伝子多様性データ群の遺伝子座位間の連鎖不平衡を、目視により一目で理解することができるようにする方法を提供することにある。

また、本発明の更なる目的は、遺伝子座位間の連鎖不平衡を、少ないコンピュータ資源で高速に算出することにある。

この発明の第１の主要な側面によれば、コンピュータシステムに、２以上の遺伝子多様性データ群の各遺伝子座における遺伝子不衡平を演算させその結果をディスプレイモニター上に比較可能に表示させるためのコンピュータソフトウエアプログラム製品であって、この製品は、記憶媒体と、この記憶媒体に格納されコンピュータシステムを動作させるための以下の指令を含む：任意の２つの遺伝子多様性データ群の各遺伝子座位の連鎖不平衡値を、その値の大きさに応じた彩度、明度、濃度を有する異なる第１、第２の色にそれぞれ変換し出力する色出力指令と；第１、第２の色を前記第１、第２の遺伝子多様性データ群間で比較可能なように前記ディスプレイモニター上に表示させる比較表示指令。ここで、前記表示指令は、前記コンピュータシステムに、各遺伝子座位の前記第１、第２の色を互いに混合させて混合色を生成させ、この混合色の配列

を第１、第２のデータ群間の連鎖不衡平値比較結果として前記ディスプレイモニター上に表示させるものであることが好ましい。

このような構成によれば、例えば、遺伝子多様性データ群のケースとコントロールデータ群の連鎖不平衡値をマトリクス状に配置し、おのおの別の色（色相の異なる色）を与え、かつ各連鎖不平衡値をその値に応じた濃さ等で表示することができる。また、このような構成によれば、比較データ群間の連鎖不平衡値の差を、混色された色やその濃さ等によって一目で認識することができる。なお、前記色はグレーカラー等の無彩色であっても良い。

また、この発明の１の実施形態によれば、この製品は、入力された前記第１、第２の遺伝子多様性データ群に基づいて、各データ群の各遺伝子座位の連鎖不平衡値を算出する連鎖不平衡値算出指令をさらに含む。ここで、この製品は、前記遺伝子多様性データ群の処理対象となる遺伝子座位の数を絞り込むための指令をさらに有することが好ましい。また、前記遺伝子座位を絞り込むための指令は、１つ又は２以上の遺伝子座位の情報エントロピーを求める手順と、上記情報エントロピーを比較して処理対象となる遺伝子座位を決定する手順とを有するものであることがさらに望ましい。１の実施形態によれば、前記情報エントロピーは、遺伝子座位のメジャーアレルに対するマイナーアレルの頻度に関する情報エントロピーであって、すべてのアレルの組合せとその頻度を用いて与えられるものである。

このような構成によれば、連鎖不平衡値算出の処理対象となる遺伝さ座位の数を、連鎖不平衡値の算出精度を落とさずに効果的に減らすことができる。なお、前記で求めた情報エントロピーの値を前記連鎖不平衡値として用いることもでき、この場合には、さらに高速で演算処理を行える。

この発明の第２の側面によれば、コンピュータシステムに、２以上の遺伝子多様性データ群の各遺伝子座における遺伝子不衡平を演算させるためのコンピュータソフトウエアプログラム製品であって、この製品は、

記憶媒体と、この記憶媒体に格納された以下の指令を含む：前記コンピュータシステムに、任意の遺伝子多様性データ群のデータを読み込む指令と；前記遺伝子多様性データ群中の任意の１又は２以上の各遺伝子座位の情報エントロピーを算出する指令と；上記情報エントロピーの値を比較して前記処理対象とする遺伝子座位を決定する手順と；前記遺伝子多様性データ群の前記処理対象となる遺伝子座位間の連鎖不平衡値を算出しコンピュータシステム上に表示するために出力する指令。ここで、前記情報エントロピーは、遺伝子座位のメジャーアレルに対するマイナーアレルの頻度に関する情報エントロピーであって、すべてのアレルの組合せとその頻度を用いて与えられるものであることが好ましい。

この発明の第３の側面によれば、コンピュータシステムに、２以上の遺伝子多様性データ群の各遺伝子座における遺伝子不衡平を演算させその結果をディスプレイモニター上に比較可能に表示させるための方法であって、任意の２つの遺伝子多様性データ群の各遺伝子座位の連鎖不平衡値を演算する工程と、前記で求めた連鎖不平衡値を、その大きさに応じた彩度、明度、濃度を有する異なる第１、第２の色にそれぞれ変換し出力する色出力工程と、第１、第２の色を前記第１、第２の遺伝子多様性データ群間で比較可能なように前記ディスプレイモニター上に表示させる比較表示工程とを有する方法が提供される。

この発明の他の特徴及び効果は、以下の発明の最良の実施形態の項に記載された好ましい実施形態と図面とを参照することによって、当業者に容易に理解することができる。

BRIEF DESCRIPTION OF THE DRAWINGS

図１は、本発明の一実施例にかかるシステム構成を説明するための概略構成図。

図２Ａ～図２Ｃは、入力データと、ケース･コントロール群の連鎖不平衡値を算出した例を示すための図。

図３は、色変換手順の構成を示す図。

図４は、第１の実施形態にかかる処理手順を示すフローチャート。

図５は、ケースとコントロール群の連鎖不平衡値を示す画面表示例。

図６は、ケースとコントロール群の連鎖不平衡値を加色混色処理した結果示すグラフィック表示例。

図７は、ケースとコントロール群の連鎖不平衡値を差分処理した結果示すグラフィック表示例。

図８は、別の実施形態にかかる処理手順を示すフローチャート。

図９は、更なる別の実施形態にかかる処理手順を示すフローチャート。

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

以下、本発明の一実施形態を添付図面を参照して説明する。

図１は、この実施形態に係るコンピュータソフトウエアをインストールしたシステムを説明するための全体構成図である。

このシステムは、ＣＰＵ１、ＲＡＭ２及び入出力部３が接続されてなるバス４に、プログラム格納部５及びデータ格納部６が接続されてなる。プログラム格納部５には、この発明の要旨に関連するもののみ挙げると、遺伝子多様性検査済データ群（遺伝子多様性データ）８を前記データ格納部６に格納するための遺伝子多様性検査済データ格納手順９と、入力されたデータから群別にペアワイズの分割表を作成して連鎖不平衡値を算出する連鎖不平衡値算出手順１０と、前記連鎖不平衡値をその値に対応した濃度の所定色の色データに変換するための色変換手順１１と、比較するデータ群間で対応する遺伝子座の色データ同士を混合した混合色データを生成する色混合手順１２と、比較するデータ群間で対応する遺伝子座同士の差分を取りその差分に応じた色及び濃度の色データ生成する連鎖不平衡値差分・色変換手順１３と、上記各手順で生成された色データをマトリックス上に整列してグラフィック表示を行うための出力表示手段１４とが格納されている。

これらの構成要素７～１４は、実際には、コンピュータシステムに設けられたハードディスク等の記憶媒体に、他の記憶媒体（ＣＤ－ＲＯＭ

等）を介してインストールされたデータ及びコンピュータソフトウエアプログラム、すなわちコンピュータシステムに対する指令である。そして、上記構成要素７～１４は、前記 CPU１によって適宜 RAM２上に呼び出され実行されることで、この発明の構成要件としての機能を奏するようになっている。また、前記入出力部３には、ディスプレイモニター１５が接続されており、前記出力表示手順１４からの出力はこのモニター１５上にグラフィカルに表示されるようになっている。

以下、これらの構成要素のさらに詳しい構成及び機能をその動作と共に説明する。

まず、前記遺伝子多様性検査データ群格納手順９が前記ＲＡＭ２上に呼び出されて実行され、遺伝子多様性検査済みデータ８群が前記データ格納部６に格納される。図２Ａに一塩基多型（図中では SNP と表示）の場合の入力データの例を示す。同図はヒトの２倍体の一塩基多型を検査した結果の例である。本データはメジャーアレルのホモを「１」に、マイナーアレルのホモを「３」に、メジャーアレルとマイナーアレルのヘテロを「２」としている。ここでメジャーアレルとは一般的に最も多い多型を意味する。また、マイナーアレルは対立遺伝子の一つで、数の相対的に少ない多型を意味する。２倍体の検査結果であるから、メジャーアレルまたはマイナーアレルいずれか一方の２つを持っている場合はホモ、それぞれ一つが混在している場合はヘテロと呼ばれる。図中「群」と言うカラム１９は「０」がケース（罹患）、「１」がコントロール（健常者）を意味する。

次に、連鎖不平衡値算出手順１０が実行され、前記遺伝子多様性検査データ群８の各遺伝子座位の連鎖不衡平値が算出される。このためには、まず前記遺伝子多様性検査データ群が前記データ格納部から呼び出されてＲＡＭ２上にコピーされる。そして、「０」のケース群と「１」のコントロール群にデータを分類し、それぞれの群に対してペアワイズな全ての組み合わせで各遺伝子座位の２×２分割表を作成する。この分割表を元に、指定に応じて D、D'、r^2、AIC 等の連鎖不平衡値を算出する。

図２Ｂ、Ｃは、前記連鎖不平衡値 r^2 を算出した例である。図２Ｂは、ケースデータ群の連鎖不平衡値の分割表であり、図２Ｃはコントロールデータ群の連鎖不平衡値の分割表である。なお、同じ遺伝座位同士は連鎖不平衡が定義されないため、空欄である（完全に連鎖していると定義することも可能）。また、この例では、全くの対象行列となるため、上三角行列だけを示し、下三角行列については表示を省略している。

r^2 では０に近い値の場合は両座位にはあまり連鎖がないことを意味する。１に近い場合は強い連鎖を持つことを意味する。したがって、図２Ｂ、Ｃの例では、SNP1 と SNP3 が強い連鎖を持ち、SNP2 と SNP4 が強い連鎖を持っていることが分かる。連鎖不平衡の計算結果としてケースデータの連鎖不平衡値とコントロールデータの連鎖不平衡値を各対応するセルごとに比較することにより、両群の連鎖の度合いの違いがある部分を見つけることができる。たとえば図２Ｂ、Ｃの例では SNP4 の列が若干違った値をとっており、ケース群とコントロール群のデータに差があることを示している。

次に、色変換手順１１が実行され、前記で求めた各連鎖不平衡値に所定の色が割り当てられる。色の割り当てがなされたら、前記出力表示手順１４が実行され、割り当てられた色が前記変換前の不平衡値と置き換えられ、前記ディスプレイモニタ１５上にマトリックス状に並べられて表示される。

この実施形態においては、前記色変換手順において決定される色は、色相（H:0～２５５）、彩度（S:0～２５５）及び明度（B:0～２５５）で表される（HSB 方式）。このため、前記色変換手順１１は、図３に示すように、色相決定手順１７と、彩度・明度決定手順１８とから構成される。

図４は、この色変換手順１１及び出力表示手順１４による処理フローである。

まず、算出されたケース群若しくはコントロール群の遺伝子座位間のペアワイズ連鎖不平衡値をメモリから取り出し（ステップＳ１）、最初の

セルから順に処理を開始する（ステップＳ２）。

そして、当該セルについて、あらかじめ設定してある色決定方法に従って色相、彩度及び明度を算出する（ステップＳ３）。すなわち、前記色相決定手順１７が、コントロール群若しくはケース群のそれぞれに割り付ける色相を所定のアルゴリズムに基いて決定する。このアルゴリズムは、比較するデータ群の数に応じて後で混色し易い色が決定されるものであるが、この実施形態では、例えばコントロール群には赤（０）、ケース群には緑(８５)が割り当てられるようにプログラミングされている。

次に、前記彩度・明度決定手順１８が、連鎖不平衡値である０．０～１．０を、その値に応じて２５６階調（０～２５５の値）の彩度及び明度に割り当て、連鎖不平衡値が高くなればなるほど同じ色相で「濃い」色になるように決定する（ステップＳ４）。

そして、出力表示手順１４は、上記ディスプレイモニタ１５上に表を描画し、各セルの連鎖平衡値を変換した色で置き換える形の画像表示を行う（ステップＳ５、ステップＳ６）。この実施形態では、上記ＨＳＢで示される色データをＲＧＢに変換して表示するようになっている。当該セルについて以上の処理が終了したならば、全てのセルについて処理が終了かを判定し（ステップＳ７）、終了していなければ上記ステップＳ３～Ｓ６の処理を繰り返す。

図５は、このようにして求められたケース群のマトリックス２１と、コントロール群のマトリックス２２とを示すモニター画面である。実際には色で表示されるが、この図５においては図示の便宜上、文字でその色を表している。この図５の画面においても、コントロール群とケース群の連鎖不平衡値が視覚的に比較可能であるが、この実施形態では、連鎖不平衡度を一目でわかるようにするため、「混色表示」及び「差分表示」を上記画面のメニューボタン２３、２４で選択できるようになっている。

混色表示を選択した場合には、前記色混合手順が実行される。

混色処理手順では、各対応するセルの描画色のＲＧＢ値の使用して上記コントロール群とケース群のペアワイズの連鎖平衡値を表す色を加色

混色により生成し、前記表示手順により混色後の色をディスプレイモニタ上に画像表示する。１つのセルについて以上の処理が終了したならば次のセルの計算に移り、全てのセルについて処理が終了するまで繰り返す。

図６に加色混色処理した結果をグラフィック表示した例を示す。上述したように、本実施例ではケース群のデータを緑に、コントロール群のデータを赤に割り付けている。したがって、混色処理の結果は、緑や赤のそれぞれの濃さに応じて、黄色～オレンジ色～緑色に表示されることになる。例えば、図に２５で示すセルは図２において該当するセルの値が両者とも０．１であり、緑、赤の色が薄く同レベルで混色され、黄色の薄い色となっている。２６で示すセルはケース、コントロールとも０．９であり、濃い黄色になっている。さらに２７で示すセルは、ケースが０．１で、コンロトールが０．０であるため薄い緑である。２８で示すセルは、ケースが０．９でコントロールが１．０であり、濃い黄色であるがわずかに赤が強いため、オレンジにやや近い値となっている。これらの処理は、具体的には、前記混色処理手順１２において、混色する２つの色間のＲ値、Ｇ値、Ｂ値の平均値を求めることで行う。

このように、ケース、コントロール群に割り当てられた色をそのまま重ねて混合して表示することにより、色の偏りがある場合はそこに連鎖不平衡の差があることを全体的に見て一目で認識することができる。

このように本実施例によれば、ケース群とコントロール群の連鎖不平衡の違いを容易に見つけられるような表示方法が可能である。

なお、本発明は上記の一実施形態に限定されるものではない。

例えば、上記一実施形態では、ケース群とコントロール群の２つを比較したが、これに限定されるものではない。別の特長による集計を適用して連鎖不平衡を求め、それらの違いを表示することも可能である。３つ以上の群を持つ場合はそれぞれ基準の群に対して差を求め、それぞれ別の色相を割り当てて表示することで３群以上の比較表示も可能である。

また、上記で色を混合することにより連鎖平衡値の差を示したが、予

め連鎖平衡値間の差を求めておいてその差分に応じて色を決定するようにしても良い。この場合はケース群の連鎖不平衡値とコントロール値を基準として連鎖不平衡値の差を求め、－１．０～０の負の値の場合は青色に、０～１．０の正の値には赤をそれぞれ絶対値にあわせて濃くするように割り当てる。

図７にこの差分表示の例を示す。この図では、ケース群とコントロール群のデータの差を取り、差異のある場所のみを表示している。セル３５は､コントロールを基準にしてケースが０．１だけ大きい場合である。大きい場合は赤に割り付けている。また、逆にコントロール群の値よりもケース群の値が小さい場合は青に割り付けてある。すなわち、－１．０～０未満は青に、０～１．０は赤に割り当てる。また両者とも絶対値が大きいほど濃い色に割り当てている。差分表示では、両者の差がどの座位間に存在するか一目で知ることができる。

なお、本実施例では、赤、青などの色を用いているがグレイスケールや、他の模様を用いことも可能である。また、一塩基多型データで説明しているが、マイクロサテライトなどのデータであってもペアワイズの分割表を作成し、その独立性の検定を行い、カイ二乗値や、そこから求められるＰ値を用いて同様に連鎖不平衡の代わりにして、そのまま画像として表示することも可能である。

また、文献 K.Shimo-onodaetal : Akaike's information criterion for a measure of linkaged isequilibrium, Journal of Human Genetics, Vol.47 Issue 12 (2002) pp649-655 に示すようにAICの独立もモデルと、従属モデルを定義してその差をとった連鎖不平衡値を利用することも可能である。カイ二乗値やAICによる連鎖不平衡の値を用いる場合にはその値の範囲が０以上の広い範囲に及ぶため、連鎖不平衡値を実際に求めた値の最大値を探索し、その最大値に対して各色をマッピングすることで、同様に視覚的に分かりやすいグラフィック表示を行うことができる。

また、色は、他の表示形式、例えばＲＧＢやＣＭＹＫによるものであっても良い。また、上記ＨＳＢ式で色を決定した後、その色をＲＧＢに

変換して処理するようにしても良い。

さらに、上記一実施形態においては、加色混色手順において、図５に示すように、まず、コントロール群とケース群について別の色を使ってカラー表示しておいてから、各セル同士の色を混色して図６に示すように混色表示を生成するようにしたが、これに限定されるものではない。図５に示すような表示を生成せずに入力データから直接図６に示す混色表示を生成するようにしても良い。

図８は、この場合の処理フローチャートを示すものである。

この図において、ステップＳ１において、混色表示対象となるセルについて、コントロール群とケース群のデータを呼び出す。ついで、当該セルについて、各コントロール群とケース群に割り当てる色相（それぞれ赤及び緑）を決定し、かつ、連鎖不平衡値の大きさに応じて色の濃さを決定する（ステップＳ２～Ｓ４）。

上記一実施形態では、ここで、コントロール群とケース群についてそれぞれグラフィックス表示を行っていたが、この例では、そのような表示を行わず、混合色を決定する（ステップＳ９）。そして、この混合色をモニター上に表示する。そして、上記セルを全てのセルについて実行する（ステップＳ１０）。

このような方法によっても、上記一実施例と同様の表示を得ることができる。

また、上記一実施形態では、前記遺伝子多様性検査データ群の全ての遺伝子座位の連鎖不平衡値を算出するようにしたが、これに限定されるものではなく、１又はそれ以上の遺伝子座位を抽出して連鎖不平値を求めるようにしても良い。一般に、１つの検査データに含まれる遺伝子座位をＮ個とすると、このうち例えば１０％を分析するのみで、分析結果の略６０％がカバーできると考えられている。したがって、そのような遺伝子座位のみを取り出して分析するようにすれば、非常に少ない計算量でそれ以上の効果を得ることが出来る。

以下、そのような遺伝子座位の抽出方法（遺伝子座位を絞り込むため

の指令手順）として、座位毎のマイナーアレルの頻度情報に着目し、その情報エントロピーを利用して特定の座位を抽出する例を図９に示すフローチャートを参照して説明する。

ここで、座位毎のマイナーアレルの頻度情報に着目して行うことが好ましいのは、同じ大きさの連鎖不平衡のものであれば、マイナーアレルの頻度がある程度高いもの同士を比較した方が、疾病に関与する遺伝子を特定しやすいからである。これは比較的少人数でマイナーアレルを持つ人を集めることができることによる。

マイナーアレルの頻度が高い遺伝子座位を採用するために、ここでは、メジャーアレルとマイナーアレルの頻度が拮抗している遺伝子座位を特定する。このための手法として、ケースデータ群の座位ごとの情報エントロピーを求めて比較する方法をとる。この情報エントロピーは、メジャーアレルとマイナーアレルの頻度をそれぞれ p, q(0<p or q<1 で p + q = 1)とすると、次式で与えられる。

情報エントロピー＝p・log2(1/p) + q・log2(1/q)

ここで、log2()は、２を底とする対数である。このようにして求められた情報エントロピーは、それぞれの遺伝子座位のアレル頻度の拮抗の度合いを明確に表す数値となり、ここでは、この数値が最も高い遺伝子座位をまず選択し、第１の遺伝子座位とする（ステップＳ１１～Ｓ１４）。

次に、この第１の遺伝子座位と組み合わせた場合に情報エントロピーが最大となる第２の遺伝子座位を選択する。この場合の情報エントロピーを求めるには、２ｘ２の分割表を用いて、まず、頻度が以下のように集計される。

第一遺伝子座位-第二遺伝子座位	頻度
1-1	p11
1-3	p13
3-1	p31
3-3	p33

この場合の情報エントロピーは次式となる。

情報エントロピー＝p11・log2(1/p11) + p13・log2(1/p13)
+ p31・log2(1/p31) + p33・log2(1/p33)

このようにして第１の遺伝子座位との組合せで情報エントロピーが最大になる遺伝子座位を決定し、これを第２の遺伝子座位として選択する（ステップＳ１４，Ｓ１５）。

この手法の利点は、ペアワイズのみでなく複数の組合せに適用できる点である。３つの組合せの場合、そのすべての組合せについて頻度を求める。例えば単一塩基多型で対立アレルが２の場合は３箇所の座位では、p111, p113, p131, p133, p311, p313, p331, p333 の８個の組合せの情報エントロピーを次式の通り計算することができる。

３座位の情報エントロピー＝p111・log2(1/p111) + p113・log2(1/p113)
+ p131・log2(1/p131) + p133・log2(1/p133)
+ p311・log2(1/p311) + p313・log2(1/p313)
+ p331・log2(1/p331) + p333・log2(1/p333)

前記ペアワイズで決定した第１、第２の遺伝子座位に対して、残りの任意の座位を第３の座位候補として組合せながら上記の情報エントロピーを算出する。その結果から情報エントロピーの最も大きなものを、第３の遺伝子座位として決定する。以下同様に第４以降の候補を追加することで、複数存在する多型の中から意味ある組合せを有効な順に決定していくことが可能である。さらに一般化して記載すると、各アレルの組み合わせのパターンがＮ種類存在し、それぞれが A1, A2, A3, ..., AN とする。また、それぞれのパターンの頻度が pA1、pA2, ..., pAN とする。ここで、pA1 + pA2 + ... + pAN = 1、0≦pA1, pA2, ..., pAN ≦1 である。これらを用いて情報エントロピーH は、次式となる。

H = pA1・log2(1/pA1) + pA2・log2(1/pA2) + ...
+ pAN・log2(1/pAN)

この遺伝子座位の抽出は、例えば、抽出した遺伝子座位の数が指定された数若しくは全体の個数に対する所定の割合になるまで繰り返される。この数は、ユーザが指定したものを用いるようにしても良いし、ユーザ

指定しない場合には、システム側が所定のしきい値を利用して決定するようにしても良い。この例では、データ群に含まれる遺伝子座位数がＮの場合、抽出した遺伝子座位の数が√Ｎに達するまで繰り返される（ステップＳ１６、Ｓ１７）。そして、このようにして決定した第１～第ｎの遺伝子座位を前記連鎖不平衡値を計算するデータ群として出力する（ステップＳ１８）

このようにして抽出された遺伝子座位群のみを使用する場合には、連鎖不平衡値をすべての組合せについて算するのではないため、最適な解が得られるとは限らないが、非常に手軽な計算で有効な遺伝子多型座位を絞り込むことが可能となる。

また、遺伝子座位の数を絞りこむのに、各遺伝子座位についてのマイナーアレルの頻度をコントロール群とケース群との間で比較し、その差の大きいものを抽出するようにしても良い。

また、次式のようにケース群とコントロール群の情報エントロピー間の差異と、両者の平均情報エントロピーを求め、それらの積を次式でしめされるように良さの指数にすることとも可能である。

良さの指数＝ケース・コントロールの情報エントロピーの差×ペアワイズの平均情報エントロピー

または、単にケース・コントロール群間の情報エントロピーの差の大きなもの上位Ｎ個の中から、平均情報エントロピーの大きなものを採用するといった発見的な手法を採用することも可能である。

さらに、上記情報エントロピーの値自体を連鎖不平衡値として用いて図４、図８の所定を行うようにしても良い。

What is claimed is:

１．　コンピュータシステムに、２以上の遺伝子多様性データ群の各遺伝子座における遺伝子不衡平を演算させその結果をディスプレイモニター上に比較可能に表示させるためのコンピュータソフトウエアプログラム製品であって、この製品は、記憶媒体と、この記憶媒体に格納されコンピュータシステムを動作させるための以下の指令を含む：

任意の２つの遺伝子多様性データ群の各遺伝子座位の連鎖不平衡値を、その値の大きさに応じた彩度、明度、濃度を有する異なる第１、第２の色にそれぞれ変換し出力する色出力指令と、

第１、第２の色を前記第１、第２の遺伝子多様性データ群間で比較可能なように前記ディスプレイモニター上に表示させる比較表示指令。

２.請求項１記載のコンピュータソフトウエアプログラム製品において、

前記表示指令は、前記コンピュータシステムに、各遺伝子座位の前記第１、第２の色を互いに混合させて混合色を生成させ、この混合色の配列を第１、第２のデータ群間の連鎖不衡平値比較結果として前記ディスプレイモニター上に表示させるものである

コンピュータソフトウエアプログラム製品。

３.請求項１記載のコンピュータソフトウエアプログラム製品において、この製品は、

入力された前記第１、第２の遺伝子多様性データ群に基づいて、各データ群の各遺伝子座位の連鎖不平衡値を算出する連鎖不平衡値算出指令をさらに含む。

４.請求項３記載のコンピュータソフトウエアプログラム製品において、この製品は、

前記遺伝子多様性データ群の処理対象となる遺伝子座位の数を絞り込むための指令をさらに有する。

５.請求項４記載のコンピュータソフトウエアプログラム製品において、

前記遺伝子座位を絞り込むための指令は、

１つ又は２以上の遺伝子座位の情報エントロピーを求める手順と、

上記情報エントロピーを比較して処理対象となる遺伝子座位を決定する手順と

を有するものである。

６．請求項５記載のコンピュータソフトウエアプログラム製品において、

前記情報エントロピーは、遺伝子座位のメジャーアレルに対するマイナーアレルの頻度に関する情報エントロピーであって、すべてのアレルの組合せとその頻度を用いて与えられるものである。

７．請求項５記載のコンピュータソフトウエアプログラム製品において、この製品は、

前記で求めた情報エントロピーの値を前記連鎖不平衡値として用いるものである。

８．コンピュータシステムに、２以上の遺伝子多様性データ群の各遺伝子座における遺伝子不衡平を演算させるためのコンピュータソフトウエアプログラム製品であって、この製品は、記憶媒体と、この記憶媒体に格納された以下の指令を含む：

前記コンピュータシステムに、任意の遺伝子多様性データ群のデータを読み込む指令と、

前記遺伝子多様性データ群中の任意の１又は２以上の各遺伝子座位の情報エントロピーを算出する指令と、

上記情報エントロピーの値を比較して前記処理対象とする遺伝子座位を決定する手順と、

前記遺伝子多様性データ群の前記処理対象となる遺伝子座位間の連鎖不平衡値を算出しコンピュータシステム上に表示するために出力する指令。

９．請求項８記載のコンピュータソフトウエアプログラム製品において、

前記情報エントロピーは、遺伝子座位のメジャーアレルに対するマイナーアレルの頻度に関する情報エントロピーであって、すべてのアレルの組合せとその頻度を用いて与えられるものである。

１０．　コンピュータシステムに、２以上の遺伝子多様性データ群の各

遺伝子座における遺伝子不衡平を演算させその結果をディスプレイモニター上に比較可能に表示させるための方法であって、

任意の２つの遺伝子多様性データ群の各遺伝子座位の連鎖不平衡値を演算する工程と、

前記で求めた連鎖不平衡値を、その大きさに応じた彩度、明度、濃度を有する異なる第１、第２の色にそれぞれ変換し出力する色出力工程と、

第１、第２の色を前記第１、第２の遺伝子多様性データ群間で比較可能なように前記ディスプレイモニター上に表示させる比較表示工程と

を有する方法。

１１．請求項１０記載の方法において、

前記表示工程は、各遺伝子座位の前記第１、第２の色を互いに混合させて混合色を生成させ、この混合色の配列を第１、第２のデータ群間の連鎖不衡平値比較結果として前記ディスプレイモニター上に表示させるものである方法。

１２．請求項１０記載の方法において、

入力された前記第１、第２の遺伝子多様性データ群に基づいて、各データ群の各遺伝子座位の連鎖不平衡値を算出する連鎖不平衡値算出工程をさらに含む方法。

１３．請求項１２記載の方法において、

前記遺伝子多様性データ群の処理対象となる遺伝子座位の数を絞り込む工程をさらに有する方法。

１４．請求項１３記載のコンピュータソフトウエアプログラム製品において、

前記遺伝子座位を絞り込む工程は、

１つ又は２以上の遺伝子座位の情報エントロピーを求める工程と、

上記情報エントロピーを比較して処理対象となる遺伝子座位を決定する工程と

を有するものである。

１５．請求項１４記載の方法において、

前記で求めた情報エントロピーの値を前記連鎖不平衡値として用いるものである方法。

１６．　コンピュータシステムに、遺伝子多様性データ群の各遺伝子座における遺伝子不衡平を演算させその結果をディスプレイモニター上に表示させるためのコンピュータプログラム製品であって、この製品は、記憶媒体と、この記憶媒体に格納された以下の指令を含む：

第１の遺伝子多様性データ群から得られた各遺伝子座位の連鎖不平衡値から、第２の遺伝子多様性データ群から得られた対応する各遺伝子座位の連鎖不平衡値を差し引かせ、その値を出力させる差し引き値出力指令と、

前記差し引き値に対応する色を生成させ、この色の配列を第１、第２のデータ群間の連鎖不衡平値比較結果として前記ディスプレイモニター上に表示させる連鎖不衡平値比較結果表示指令。

１７．　コンピュータシステムに、遺伝子多様性データ群の各遺伝子座における遺伝子不衡平を演算させその結果をディスプレイモニター上に表示させるための方法であって、

第１の遺伝子多様性データ群から得られた各遺伝子座位の連鎖不平衡値から、第２の遺伝子多様性データ群から得られた対応する各遺伝子座位の連鎖不平衡値を差し引かせ、その値を出力させる差し引き値出力工程と、

前記差し引き値に対応する色を生成させ、この色の配列を第１、第２のデータ群間の連鎖不衡平値比較結果として前記ディスプレイモニター上に表示させる連鎖不衡平値比較結果表示工程と

を有する方法。

ABSTRACT OF THE DISCLOSURE

コンピュータシステムに、２以上の遺伝子多様性データ群の各遺伝子座における遺伝子不衡平を演算させその結果をディスプレイモニター上に比較可能に表示させるためのコンピュータソフトウエアプログラム製品であって、この製品は、記憶媒体と、この記憶媒体に格納されコンピュータシステムを動作させるための以下の指令を含む：

任意の２つの遺伝子多様性データ群の各遺伝子座位の連鎖不平衡値を、その値の大きさに応じた彩度、明度、濃度を有する異なる第１、第２の色にそれぞれ変換し出力する色出力指令と、

第１、第２の色を前記第１、第２の遺伝子多様性データ群間で比較可能なように前記ディスプレイモニター上に表示させる比較表示指令。